

Global expansion of *Mycobacterium tuberculosis* Lineage 4 shaped by colonial migration and local adaptation

Ola B Brynildsrud,¹ Caitlin S Pepperell,^{2,3} Philip Suffys,⁴ Louis Grandjean,⁵
Johana Monteserin,^{6,7} Nadia Debech,¹ Jon Bohlin,¹ Kristian Alfsnes,¹
John Pettersson,¹ Ingerid Kirkeleite,¹ Fatima Fandinho,⁸ Marcia Aparecida da Silva,⁸
Joao Perdigao,⁹ Isabel Portugal,⁹ Miguel Viveiros,¹⁰ Taane Clark,^{11,12}
Maxine Caws,^{13,14} Sarah Dunstan,¹⁵ Phan Vuong Khac Thai,¹⁶ Beatriz Lopez,⁶
Viviana Ritacco,^{6,7} Andrew Kitchen,¹⁷ Tyler S Brown,¹⁸ Dick van Soolingen,¹⁹
Mary B O'Neill,³ Kathryn E Holt,^{20,21} Edward J Feil,²²
Barun Mathema,²³ Francois Balloux,²⁴ Vegard Eldholm^{1*}

¹ Infectious Diseases and Environmental Health, Norwegian Institute of Public Health, Lovisengergata 8, 045

² Division of Infectious Disease, Department of Medicine, School of Medicine and Public Health, University of

³ Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of

⁴ LABMAM - IOC - FIOCRUZ, Av Brasil 4365, Manguinhos 21040-360, CP 926, Rio de Janeiro, Brazil

⁵ Department of Paediatric Infectious Diseases, Imperial College London, W2 1NY, London, UK

⁶ Instituto Nacional de Enfermedades Infecciosas, ANLIS Carlos Malbran, Buenos Aires, Argentina

⁷ Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Buenos Aires, Argentina

⁸ Laboratorio de Bacteriologia da Tuberculose Centro de Referência Professor Helio Fraga-Jacarepagu (Estrada

⁹ Instituto de Investigação do Medicamento, Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal

¹⁰ Unidade de Microbiologia Médica, Global Health and Tropical Medicine, Instituto de Higiene e Medicina T

¹¹ Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, W

¹² Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London

¹³ Liverpool School of Tropical medicine, Department of Clinical Sciences, Liverpool, UK

¹⁴ Birat-Nepal Medical Trust, Lazimpat, Kathmandu, Nepal

¹⁵ Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia

¹⁶ Pham Ngoc Thach Hospital for TB and Lung Diseases, Ho Chi Minh City, Vietnam

¹⁷ Department of Anthropology, University of Iowa, Iowa City, IA 52242, USA

¹⁸ Division of Infectious Diseases, Massachusetts General Hospital, Boston, USA

¹⁹ Center for Infectious Disease Research, Diagnostics and Perinatal Screening, National Institute for Public Health, Stockholm, Sweden

²⁰ Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Australia

²¹ Bio21 Institute, University of Melbourne, Melbourne, Australia

²² Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

²³ Mailman School of Public Health, Columbia University, 722 W 168th St, New York, NY 10032, USA

²⁴ UCL Genetics Institute, University College London, London WC1E 6BT, UK

^{1*}To whom correspondence should be addressed; E-mail: elve@fhi.no

Based on population genomic and phylogeographic analyses of 1669 *Mycobacterium tuberculosis M.tb* Lineage 4 (L4) genomes, we find that dispersal of L4 has been completely dominated by historical migrations out of Europe. We demonstrate an intimate temporal relationship between European colonial expansion into Africa and the Americas and the spread of L4 tuberculosis (TB). Strikingly, in the age of antibiotics, mutations conferring antimicrobial resistance (AMR) overwhelmingly emerged locally (at the level of nations), with minimal cross-border transmission of resistance. The latter finding was found to reflect the relatively recent emergence of these mutations, as a similar degree of local restriction was observed for susceptible variants emerging on comparable time-scales. The restricted international transmission of drug resistant TB suggests that containment efforts at the level of individual countries could be successful.

Introduction

Tuberculosis (TB) takes more lives than any other infectious disease. Global TB burden has declined slowly over the last decade but the rise of antimicrobial resistance (AMR) constitutes a significant obstacle to TB-control in the absence of an effective vaccine. In recent years, a number of attempts have been made to reconstruct the evolutionary history of *Mycobacterium tuberculosis* (*M.tb*) and its association with humans. One genome-based study hypothesised that *M.tb* spread out of Africa together with early humans (1), whereas a later study, employing ancient DNA samples for temporal calibration, suggested a far younger most recent common ancestor (MRCA) of extant *M.tb* 4,000-6,000 years ago (2). Among seven recognized *M.tb* lineages, lineage 4 (L4) is the most widely dispersed, affecting humans across the world. Here, relying on a collection of 1,669 L4 genomes, including hundreds of novel genomes from the Americas, we set out to reconstruct the migration history of L4 and assess the impact of migration on the spread of AMR. We find that repeated sourcing from Europe has been the main driving force for the global expansion of L4, with intense dispersal to Africa and the Americas concomitant with European colonizing efforts ca 1600-1900 CE. We also find that the rise of multidrug-resistant TB (MDR-TB) in recent decades is overwhelmingly a local phenomenon, in the sense that resistant clones have emerged repeatedly in a wide variety of locations, while migration of resistant strains seems to have played a marginal role in shaping the observed L4 AMR landscape.

Results and Discussion

In total, 1669 genomes representing clinical *M.tb* isolates from 15 countries were included in the study (Fig. 1). After down-sampling of densely sampled outbreaks (Materials and Methods), 1205 isolates remained. To get an overview of global patterns of genomic diversity and

strain distribution, we assigned each genome to a sub-lineage based on the Coll scheme (3) and mapped the sub-lineage annotations on the temporal phylogeny (Fig. 1). Generally, sub-lineages were found to be widely dispersed, but clear patterns of geographic structure are discernible, as noted by Stucki *et al* (4): L4.5 was restricted to **South East Asia (Vietnam)**, whereas L4.3 (also termed LAM) was underrepresented in the country (Fig. 2). We also observed early, distinct splits within sub-lineages 4.2 and 4.4; L4.2 consists of a **Vietnamese** cluster nested within the otherwise strictly European sub-lineage, whereas an early branching event separates L4.4 into two clades, one exclusively detected in Vietnam and the other global in its distribution.

To assess the overall L4 diversity as a function of geography, we investigated the distributions of pair-wise SNP differences in countries where sampling was deemed to be sufficiently dense and representative (Fig. 2). Russia, the Netherlands and Vietnam were characterized by the highest diversity in circulating L4 strains, both in terms of median pairwise distance and Simpson's diversity (which also takes sub-lineage distribution into account). Comparing the patterns of diversity in Malawi and Vietnam, representing **large collections from Karonga district and Ho Chi Minh City, respectively (5,6)**, the SNP-distance distributions indicate more ongoing transmission of L4 in Malawi (where pronounced peaks in the lower tail of the pairwise-distance distribution is visible) relative to Vietnam.

Next we performed analyses to formally assess the phylogeographic history of L4. We employed both discrete trait analyses (DTA) (7) and Bayesian structured coalescent approximation (BASTA) (8)). For both analyses a temporal phylogeny inferred with Beast 1.8.4 (9) was used as an input tree and sampling location used to assign samples to one of five regions: Europe, Africa, South East Asia, South America and North America (Russian isolates were assigned to Europe, since all isolates were sampled West of Ural, in accordance with UN geoschemes.)

The L4 MRCA was estimated to have existed around 1096 CE [95% HPD: 955-1231] (Supplementary Materials). Applying DTA, the geographic location of the MRCA was estimated to be Europe with high confidence (posterior probability=0.92).

BASTA initially placed the root location in South America, which in light of the estimated age of L4 is highly unlikely, and also necessitated migration events to the Old World in the centuries immediately preceding the Columbian discovery of the Americas. We thus performed individual analyses with the root location forced to be either of the three Old World continents to which our samples belonged. In these analyses, a European root location was overwhelmingly favored (Materials and Methods). The resulting migration matrices estimated by DTA and BASTA were largely concordant (Fig. S5), both methods inferring Europe as playing a pivotal role in the global dispersal of L4. The central role of Europe in the global expansion of L4 was most pronounced in the BASTA inferences, where out-of-Europe migration was found to be almost singlehandedly responsible for the current geographic range of L4 (Fig. 3A). However, "Europe" should not be interpreted in the strict sense, and probably captures interactions in the Middle ages with West Asia and North Africa as well. Thus our study does not contradict an origin around the Mediterranean as suggested in a recent study (10).

To investigate the migration history of L4 in a temporal context, we analyzed the inferred load and direction of migration over time (see Materials and Methods). As it was clear that out-of-Europe migrations were the overwhelming driver of global L4 range expansion (Fig. 3A), we illustrate migration from Europe to the other continents through time in Fig. 3B. In parallel, we analyzed migration events within the receiving continents to get a picture of the relative importance of import versus intra-continental transmission (Fig. 3C).

Our phylogeographic analyses suggest that the first waves of L4 migration out of Europe

were in an eastward direction, with the first migration to South East Asia (represented by Vietnam) estimated to the beginning of the 13th century. Here, local populations were quickly established, and internal transmission became dominant by the late 16th century (Fig. 3C). These observations fit well with the observed population structure of the Vietnamese isolates (Fig. 1) and the high diversity observed within the country (Fig. 2). Present-day Vietnam was part of a large French colony termed French Indochina from the late 19th century onwards. France was not among the sampled countries in the current study, but focusing on the Netherlands, the most proximal, sampled country, we identified a number of nodes with Vietnamese and Dutch descendants only, the first dated to 356-426 years before present (95% HPD), followed by six nodes with 95% HPDs from 114 to 269 years ago. These date ranges fit well with known French-Vietnamese interactions, starting with the arrival of the missionary Alexandre de Rhodes in the 1627 (11) followed by French military expansion from the mid 19th century and the formation of French Indochina in 1887.

The next waves of migration were directed towards Africa, with the earliest introduction among the sampled countries inferred in the present-day Republic of Congo (Congo-Brazzaville) in the 15th century and subsequent introductions to South Africa, Uganda and Malawi (Southern and Eastern Africa) from the late 17th century. The early introduction in Congo is likely a result of the relatively proximity of Congo-Brazzaville to the West African territories which were first to interact with European explorers. In contrast to what we observed for South East Asia, repeated sourcing from Europe seems to have been more important than local transmission until the 19th century (Fig. 3C). These findings closely mirror the European colonial history in Africa south of the Sahara, with early Portuguese forts and trading posts established on the Gold Coast (present day Ghana) in 1482, followed by an ever increasing European presence and influence in African coastal regions over the next centuries. Finally, this culminated in an all-out

land grab termed the "scramble for Africa" in the late 19th century, placing vast portions of the African continent under European control in the form of colonies. Our findings thus suggest an intimate relationship between European colonial expansion and the spread of L4 tuberculosis on the African continent. The main form of internal African migration in this time period was connected to the expansion of the Zulu Empire under Shaka (1816-1828) which forced fleeing tribes to migrate north and eastwards. This appears to have been less important in the spread of L4, but internal nodes with descendants in Congo, Uganda and Malawi dated to the 18th-20th century point to secondary transmission-routes through internal African migration.

The three earliest migration events to the Americas were inferred to have occurred between 1466 and 1593 to South America and between 1566 and 1658 to North America. These migration events occurred along long branches of the phylogeny, so the exact timing cannot be established. The estimates do however suggest that Europeans brought TB to South America relatively soon after the arrival of Europeans on the continent in 1492. An abrupt increase in the flow of L4 into South America is seen from the turn of the 17th century (Fig. 3). Bone pathology and lesions identified in human remains suggest that TB in the Americas might have pre-dated European contact (12), but convincing molecular evidence is limited to the identification of *M. pinnepedii* infection (a *M.tb* complex-member generally restricted to seals and sea lions) in skeletal samples from a Peruvian coastal settlement (2). As the original human colonization of the Americas pre-dates the likely age of the *M.tb* MRCA (2) by a large margin, the most parsimonious interpretation is that whilst animal strains were in circulation in some Native American human populations, *M.tb sensu stricto* was introduced to the Americas by European colonists and followed by continued influx of L4 with subsequent waves of European migrants. The near complete dominance of L4 in South America (13) also supports this notion.

Despite the massive and relentless import of TB, we find that the establishment of detectable local transmission was delayed in South America relative to Africa and South East Asia (Fig. 3), which might reflect the massive decline of native populations following European contact. Infectious diseases caused severe die offs of native populations following the arrival of Europeans (14). The toll taken by infectious disease epidemics in Native Americans is ascribed to their vulnerability to a number of pathogens introduced by Europeans, and was likely exacerbated by widespread societal collapse and famines. The near wholesale replacement of native Americans with Europeans and African slaves in many regions (14) might explain the delayed imprints of local TB transmission in our phylogeographic analyses. **A figure including migration over time to both North (relatively sparsely sampled after down-sampling of outbreaks) and South America is included in Fig. S6.**

The phylogeographic analyses also allowed us to shed light on the history of specific strains of interest. In South Africa, the KZN strain is responsible for a devastating epidemic of XDR-TB (15, 16). We find that the ancestor of the KZN strain was dispersed with Europeans to South Africa about 130 years ago (Fig. 3B). In the same period, about 100-150 years ago, there were independent introductions of a closely related strain from Europe to Latin America, providing context to the observation by Lanzas *et al.* that strains closely related to the KZN strain are driving an MDR-TB outbreak in Panama (17).

Another interesting finding concerns the RdRio clade, originally identified as a major cause of TB in Brazil (18), but later identified at moderate to high frequencies also in Portugal, the US and beyond (19, 20). Our results suggest that the RdRio clade originated in Europe around 350 years ago, followed by multiple introductions to Africa and the Americas from around 250 years ago. **To investigate this clade in more detail, we extracted the RdRio subtree from the full**

L4 phylogeny and performed an independent phylogeographic analysis. This analysis indicated Iberia as a likely origin of RdRio, followed by early expansions to Atlantic South America, Peru and South-East Africa (Fig. S7).

Furthermore, we find that the ancestors of two major MDR-outbreak strains in Argentina, M (21) and Ra (22), were both introduced to South America around 200 years ago. The L4 DS6_{Quebec} clade is common among Aboriginal populations in Ontario, Saskatchewan and Alberta as well as French Canadians in Quebec. As substantial contact between these populations was limited to the period of 1710 to 1870 (23), this provides a useful interval against which to test our temporal inferences. Our analyses places the MRCA of DS6_{Quebec} in Canada in 1788 [95% HPD: 1739-1827], well within this interval. The inferred timing of L4 migration to Africa and the Americas fits remarkably well with the known history of European colonization of the continents. A summary of the direction, intensity and timing of L4 migration is included in Fig. 3.

For clonal, non-recombining organisms such as *M.tb*, the identification of homoplastic mutations is a powerful way to identify targets of selection (24). We identified a total of 733 mutations that had emerged more than once in the L4 dataset. As expected, the top-scoring genes included a number of known AMR genes (dataset S02). Intriguingly, we also identified a handful of promoter and non-synonymous genic mutations (codons 3 and 253) in the lactate dehydrogenase gene *lldD2*, that had evolved independently >100 times. A recent study demonstrated that *lldD2* is important for *M.tb* replication within human macrophages by enabling the bacillus to utilize lactate as a carbon source (25). A screen for positive selection in a smaller dataset covering *M.tb* lineages 1-6 (26) found that the codon 3 mutation had emerged independently in lineages 1,2 and 4 whereas the codon 253 mutation had emerged repeatedly in L4 and

was present in all but a single L2 isolates. In L4, we find that *ltdD2* mutations started emerging well before the age of antibiotics (Fig. S9) and have emerged across all continents (Dataset S3), suggesting parallel local adaptation, most likely to broad changes in host ecology. Finally, we assessed the transmissibility of clades with and without *ltdD2* mutations, both in terms of number of descendants per node age and in terms of transmission across country borders (see supplementary text). These analyses suggested that there were indeed differences between the groups, and that strains harboring *ltdD2* promoter mutations carry a significant benefit in terms of transmissibility.

The most serious challenge to TB control efforts is the rise of AMR, which is threatening to reverse the moderate decrease in TB burden obtained over the last decade (27). In order to investigate the role of migration in the spread of AMR, we mapped the time and location of resistance emergence by mapping known resistance mutations on the temporal phylogeny. For clarity, we only included genes relevant for the multi- and extensive- drug resistance definitions (MDR-TB: resistance to isoniazid [INH] and rifampicin [RIF]; XDR-TB: MDR-TB with additional resistance to fluoroquinolones [FLQ] and one of the injectable drugs kanamycin, amikacin or capreomycin). As the major mutations responsible for kanamycin, amikacin and capreomycin resistance are largely overlapping, we refer to this group as KAN resistance mutations. The stacked bar chart summarizing the inferred timing of individual resistance emergence events (Fig. 4, top panel) indicates a gradual increase in AMR emergence rate from the 1960s till the late 1990s, after which a plateau is reached. In line with an earlier global study of AMR in *M.tb* (28), we find that MDR-TB has emerged repeatedly and independently across geographic regions (Fig. 1).

In fact, when extracting the geographic location of isolates descending from resistance-

nodes (i.e. inherited resistance), we did not identify a single instance of a resistant strain crossing country borders. To understand the underlying cause of this observation, we first investigated whether this reflected a decrease in transmission-fitness of resistant strains (see Materials and Methods). This analysis did not indicate any decrease in transmissibility of resistant strains relative to their susceptible counterparts (Fig. 4 bottom panel). We thus hypothesized that the failure of resistant strains to cross country borders **in our global dataset** might simply reflect the young age of these strains. Cross-country and cross-continental migration events as a function of node age and the predicted phenotypic resistance of the strain occupying the node were thus quantified. From Fig. 4 (middle panel) it is clear that cross-border migration was exceedingly rare among descendants of nodes young enough to be resistance nodes, irrespective of susceptibility profile. In fact, only a single node emerging within the age of antibiotics (post 1945) was found to have descendants that had crossed country borders in our dataset. We thus conclude that the young age of resistance-nodes, rather than decreased transmission-fitness explains the lack of observed migration of these strains.

As a result of the clonal mode of *M.tb* replication (29), efficient adaptive evolution across populations requires the parallel evolution of beneficial mutations. This is indeed the pattern we observe both for adaptive mutations in the lactate metabolism gene *lldD2* and, importantly, the emergence of multidrug-resistance across the L4 phylogeny (Fig. 1). AMR emergence within L4 tuberculosis reflects local adaptations to near-identical treatment schemes across diverse geographic contexts. There is no doubt that resistant *M.tb* strains can cross country borders, and has been observed *e.g.* in the case of resistant Lineage 2 isolates imported from Eastern to Western Europe (30), but we demonstrate that migration has played a negligible role in shaping global AMR patterns in L4. The geographic restriction of resistant strains is indeed striking, and suggests that the challenge of AMR can still be tackled efficiently at the level of individual

nations. If, however, we fail to act swiftly using the best informed interventions available, this picture might change rapidly.

References

1. I. Comas, *et al.*, *Nat Genet* **45**, 1176 (2013).
2. K. I. Bos, *et al.*, *Nature* **514**, 494 (2014).
3. F. Coll, *et al.*, *Nat Commun* **5** (2014).
4. D. Stucki, *et al.*, *Nature Genetics* **48**, 1535 (2016).
5. J. A. Guerra-Assuncao, *et al.*, *eLife* **4** (2015).
6. K. E. Holt, *et al.*, *bioRxiv* (2016).
7. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, *PLoS Comput Biol* **5**, e1000520 (2009).
8. N. De Maio, C.-H. Wu, K. M. O'Reilly, D. Wilson, *PLOS Genetics* **11**, e1005421 (2015).
9. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, *Molecular Biology and Evolution* **29**, 1969 (2012).
10. M. B. O'Neill, *et al.*, *bioRxiv* (2017).
11. O. Chapuis, *A History of Vietnam: From Hong Bang to Tu Duc* (1995).
12. H. D. Donoghue, *Microbiology Spectrum* **4** (2016).
13. V. Ritacco, *et al.*, *Mem Inst Oswaldo Cruz* **103**, 489 (2008).
14. N. D. Cook, *Born to Die - Disease and New World Conquest, 1492-1650* (Cambridge University Press, 1998).
15. K. A. Cohen, *et al.*, *PLoS Med* **12**, e1001880 (2015).

16. N. S. Shah, *et al.*, *New England Journal of Medicine* **376**, 243 (2017).
17. F. Lanzas, P. C. Karakousis, J. C. Sacchettini, T. R. Ioerger, *Journal of Clinical Microbiology* **51**, 3277 (2013).
18. L. C. O. Lazzarini, *et al.*, *Journal of Clinical Microbiology* **45**, 3891 (2007).
19. S. A. Weisenberg, *et al.*, *Infection, Genetics and Evolution* **12**, 664 (2012).
20. S. David, *et al.*, *Infection, Genetics and Evolution* **12**, 1362 (2012).
21. V. Eldholm, *et al.*, *Nature Communications* **6** (2015).
22. V. Ritacco, *et al.*, *Emerg Infect Dis* **18**, 1802 (2012).
23. C. S. Pepperell, *et al.*, *Proc Natl Acad Sci U S A* **108**, 6526 (2011).
24. M. R. Farhat, *et al.*, *Nat Genet* **45**, 1183 (2013).
25. S. Billig, *et al.*, *Scientific Reports* **7**, 6484 (2017).
26. N. S. Osório, *et al.*, *Molecular Biology and Evolution* **30**, 1326 (2013).
27. K. Dheda, *et al.*, *Lancet Respir Med* (2017).
28. A. L. Manson, *et al.*, *Nat Genet* **49**, 395 (2017).
29. V. Eldholm, F. Balloux, *Trends in Microbiology* **24**, 637 (2016).
30. N. Casali, *et al.*, *Nat Genet* **46**, 279 (2014).

Acknowledgments

We are grateful to Nicola de Maio for helpful tips with setting up the BASTA XML. We would also like to acknowledge the Norwegian Sequencing Center for their efficient and competent handling of our sequencing needs.

Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1 to S8

References (27-76) Supplementary Datasets S01 to S03

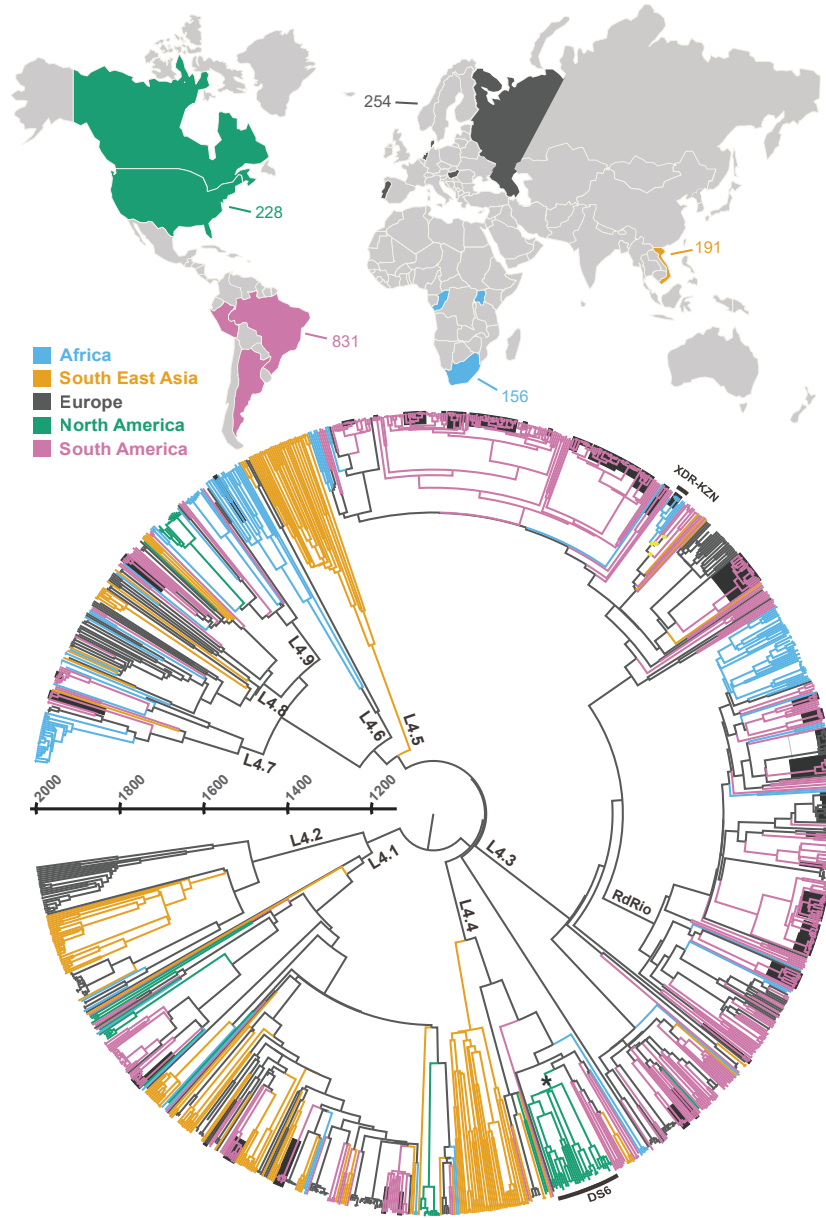


Figure 1: *Sampling overview and phylogeography of the global L4 dataset. In the map, sampled countries are colored by continent and sample sizes indicated. In the temporal phylogeny, branches are colored to match the most likely geographic location inferred using BASTA. MDR-clusters identified in the dataset are highlighted with black background shading. A large black asterisk highlights the branch leading to the DS6_{Quebec} clade, used to assess robustness of dating analyses, whereas yellow dots indicate independent introductions of the KZN ancestor to South Africa and South America. See main text for details.*

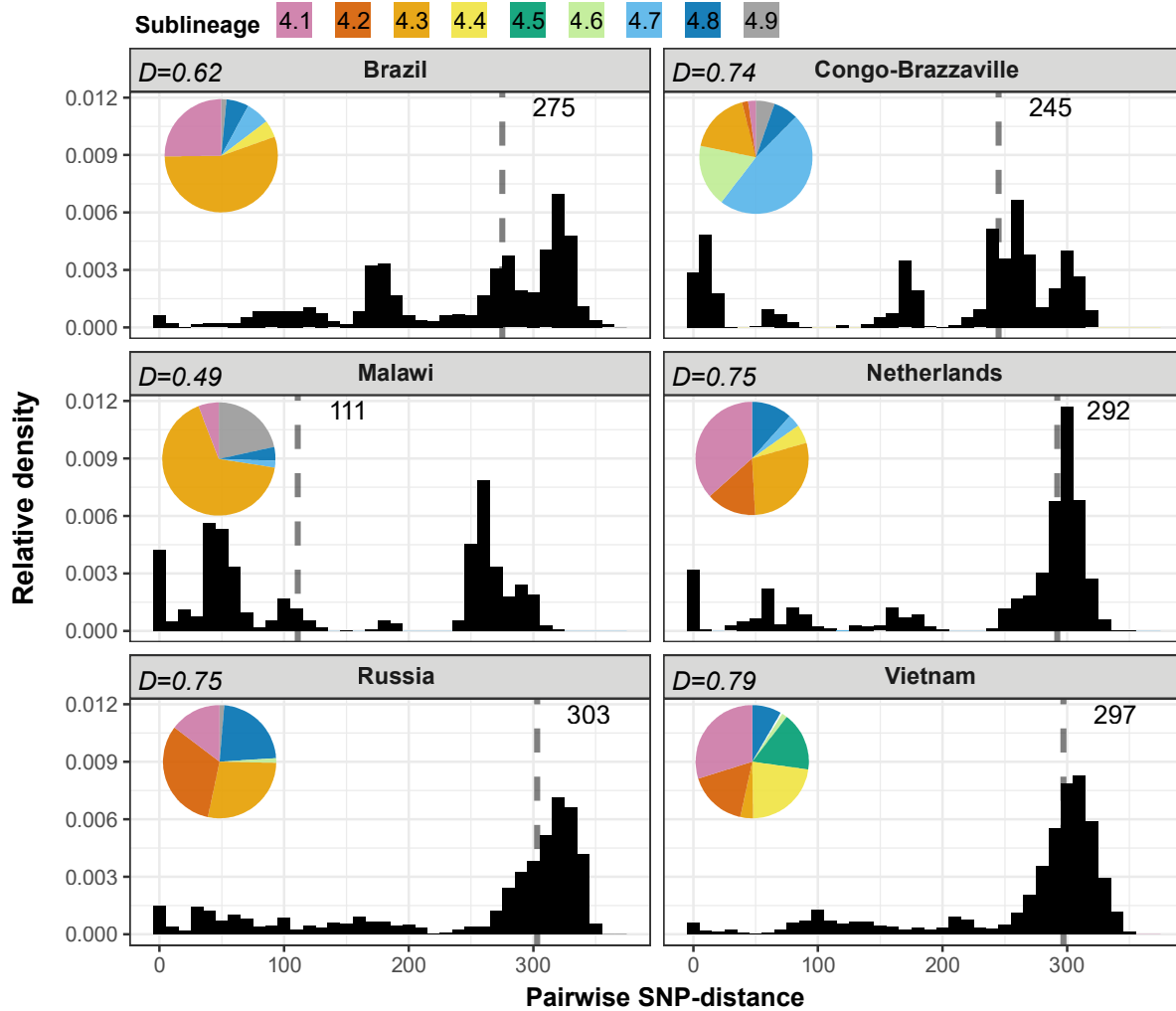


Figure 2: Within-country diversity as assessed by mean pairwise SNP-distances (only including well-sampled countries). Vertical dashed lines indicate median values. Embedded pie charts summarize sub-lineage distribution within each country. The Simpsons diversity index ($1-D$) was calculated at the level of sub-species and the estimate indicated in the top of each country panel (where higher values correspond to increased diversity).

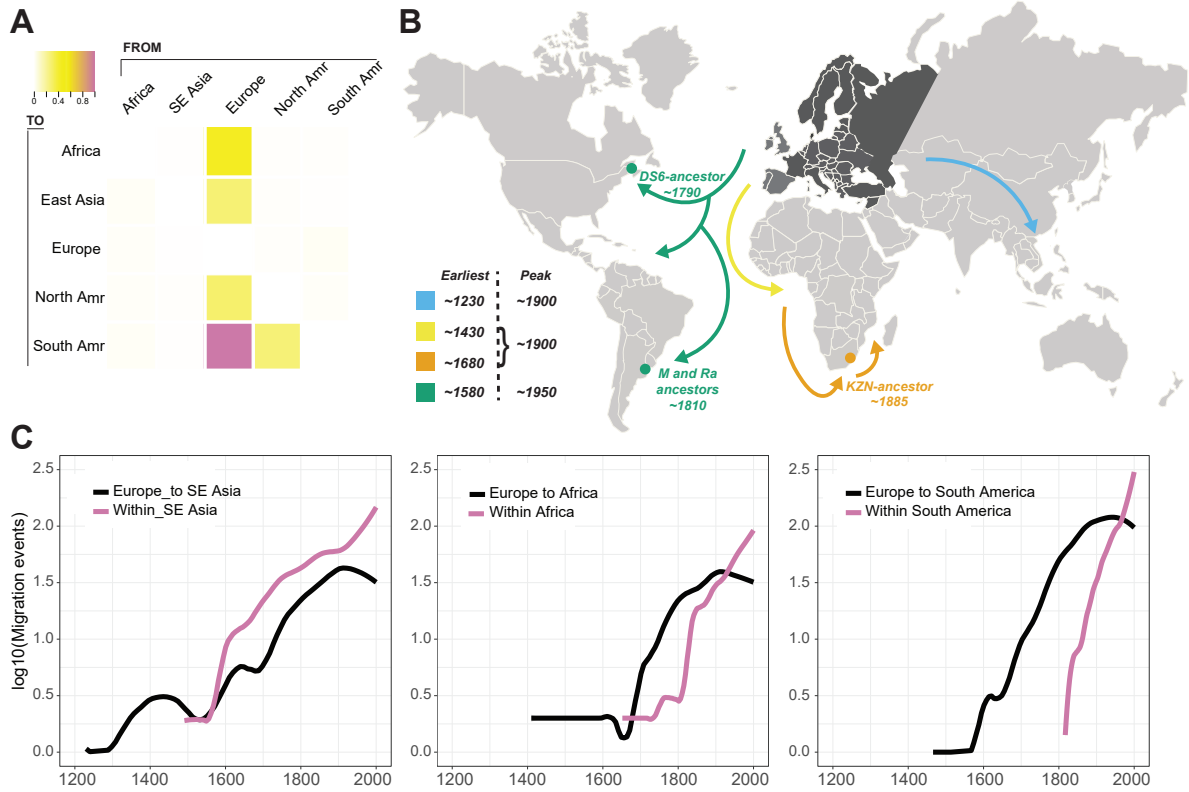


Figure 3: Lineage 4 migration. (A) Heat-map summarizing over-all migration load between continents as inferred in BASTA. (B) Temporal overview of L4 migration out of Europe. The establishment of strains of interest discussed in the text are highlighted. As the exact timing of the first American migrations was uncertain, the mean of the first three inferred migration events to each of the two subcontinents is reported as an approximation of the earliest migration events to the Americas. (C) Out-of-Europe migration to South East Asia, Africa and South America over time. The plots also show within-continent migration/transmission in the receiving continents to illustrate the relative importance of repeated L4 import on continental L4 load over time.

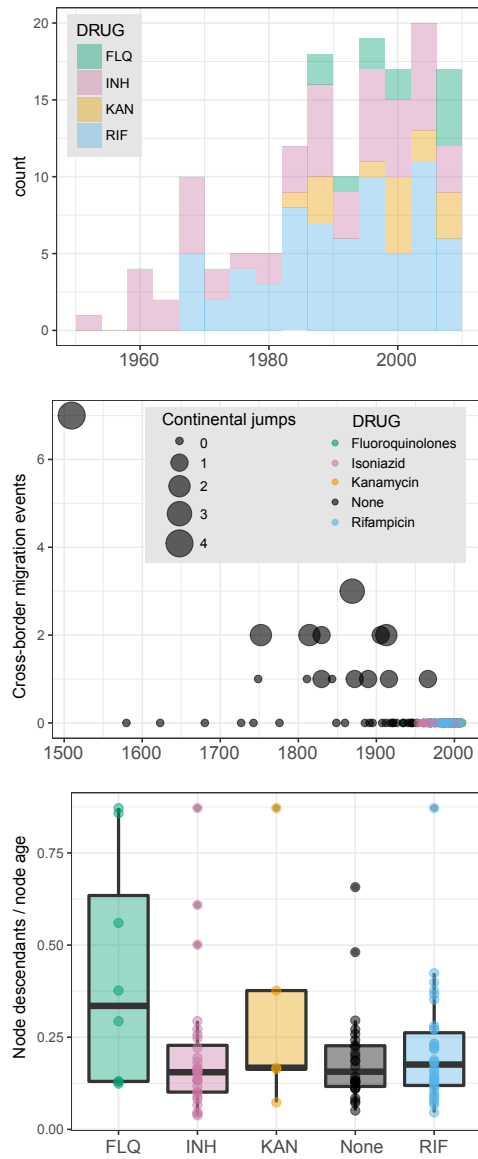


Figure 4: *Transmission of resistance. Top panel: Independent emergence of AMR over time based on the age of nodes where resistance mutations were inferred to have emerged. The middle panel illustrates inferred cross-border transmission of the descendants of susceptible and resistant ancestors as a function of node age. The size of the dots indicate the number of inferred cross-continental migration events occurring among ancestors of each node. Individual dots are colored by the drug to which they cause resistance. As a proxy for transmissibility, The bottom panel shows the number of descendants divided by node age for inferred nodes with or without resistance mutations.*

Supplementary Text

June 18, 2018

Rationale for restricting age of resistance nodes

We know of one report of *M.tb* AMR mutations having emerged before the age of antibiotics in the case of pyrazinamide (31). Mutations yielding pyrazinamide resistance were not included in the study as they are not covered by the MDR and XDR definitions. As an obligate and largely intracellular parasite, the bacillus is generally shielded from antimicrobial compounds produced by competing bacteria, and there is no doubt that AMR mutations overwhelmingly emerged in the age of antibiotics. Allowing for resistance emergence prior to clinical use of antibiotics clearly has the potential to cause artifactual results in the form of very early emergence of common resistance mutations (*e.g.* *rpoB* S450L and *katG* S315T). We thus applied the following rule when mapping AMR mutations to nodes in order to date their emergence: Resistance emergence was not allowed to occur at nodes older than the year of discovery of the drug - in such instances, the mutation was mapped to the next pair of descendant nodes.

Selection on *lldD2* mutations

Our homoplasy scan identified a number of known antimicrobial resistance (AMR) genes among the top-scoring hits (dataset S02). Strikingly, a high number of homoplastic sites were identified in and immediately upstream of the lactate dehydrogenase gene *lldD2*. In fact, the parsimony

score of *lldD2* was second only to *embB* (Dataset S02). Two non-synonymous mutations (V3I and V253M) had emerged independently 32 and 43 times respectively, whereas five mutations immediately upstream (likely in the promotor region, see dataset S3 for exact positions) had emerged a total of 40 times. The *lldD2* mutations were found to have emerged on all continents (dataset S03), and started emerging well before the age of antibiotics (Figures S2 and S9), indicating that *lldD2* mutations did not emerge in response to drug-challenge.

M.tb is an obligate human parasite, and the *lldD2* mutations thus most likely constitute adaptive response to global changes in host ecology, such as increased human population densities or possibly changes in nutrition. A recent study demonstrated that *lldD2* is important for *M.tb* replication within human macrophages by enabling the bacillus to utilize lactate as a carbon source (32). Even though the *M.tb* life cycle within humans is incompletely understood (33), macrophages play an integral role, and mutations tuning the metabolism of the bacillus within macrophages could thus be of major importance.

To test whether clones harboring *lldD2* mutations tended to spread more efficiently than their wildtype counterparts, we analyzed the transmissibility of clones with and without *lldD2* mutations, using the number of descendants per unit time as a proxy for transmission fitness (34) (Fig. S2 and supplementary materials and methods). The slope of the regression is clearly steeper for clones harboring codon 3 and promoter mutations relative to the controls and those harboring codon 253 mutations, indicating that these clones might be expanding more efficiently. ANCOVA analysis (see supplementary materials and methods) indicated a positive association between *lldD2* promotor mutations and transmissibility. (F-test: $p=0.038$). The *lldD2* mutations tending to leave more descendants (promoter and V3I mutations) were also found to have emerged more recently (the vast majority post 1900, See Figures S2 and S9),

suggesting that they have evolved in response to recently emerged selective pressures.

Fig. S1 visualizes the all identified homoplastic sites as a function of gene-wise Watterson's θ and nucleotide diversity (π).

Dating analyses in relation to earlier studies

Two earlier studies employing ancient DNA for temporal calibration estimated the TMRCA of L4 to 330 BC - 761 CE (Bos *et al.*) (35) and 40 CE - 662 CE (Kay *et al.*) (36). The Bos *et al.* study focused on *M.tb* as a whole, and L4 rate estimates will thus be affected by the overall estimated rate, when this is not constant across lineages. The ancient genomes published by Kay *et al.* are in fact the same genomes we used for temporal calibration here. In this instance there must be other reasons for the moderate deviation in estimated TMRCA. One possible source of difference could be that we excluded one out of four genotypes due to low coverage. As the mummy genomes are responsible for most of the temporal structure in our dataset, excluding one genome could affect TMRCA estimates. Bos *et al.* estimated a substitution rate for the *Mycobacterium tuberculosis* complex (MTBC) of 4.6×10^{-8} substitutions per site per year (95% HPD: 3.0×10^{-8} to 6.2×10^{-8}), whereas Kay and colleagues estimated the rate to 5.0×10^{-8} (95% HPD: 4.1×10^{-8} to 5.9×10^{-8}). Our estimated rate of 4.8×10^{-8} (95% HPD: 4.2×10^{-8} - 5.4×10^{-8}) fall in the middle of these already very similar estimates.

Patient origins and phylogeographic inferences

Samples were assigned to geographic regions based on sampling location. However, a subset of the patients are bound to be immigrants, of which some are likely to have acquired their infection prior to their arrival in the country where the isolate was ultimately sampled. If such instances are very common could potentially affect phylogeographic inferences. We thus ob-

tained information on the country of origin of individual samples from Portugal, a central player during the Age of Exploration and ensuing European colonial activities. Of 51 genomes from Portugal included in the study, 28 of the sampled patients were of European origin, 26 of which were Portuguese. Samples from two patients from the Cape Verde islands, formally part of Africa, but uninhabited before the arrival of the Portuguese, were both embedded in clusters of European isolates. Another six patients were originally from Angola (2), Brazil, Mozambique, Sao Tome e Principe or Venezuela. For the remaining Portuguese samples the country of origin of the patients were unknown. We were also able to obtain data on the country of origin of Dutch patients. A total of 34 Dutch patients had a country of origin outside of the Netherlands. This included 11 from other European countries including Turkey, 10 from Suriname, Aruba and Brazil, two from Somalia, one from Indonesia and 14 from Morocco. For the separate phylogeographic analyses of the RdRio clade, we uses the country of origin as category, rather than country of isolation. The isolates were sorted into one of eight categories (name categories). Two genomes isolated in the Netherlands were of Spanish origin, these were added together with Portuguese isolates in the Iberia category. Importantly, manual inspection of the whole 1207-taxon L4 phylogeny revealed that none of the Dutch/Portuguese isolates situated on particularly long branches or otherwise situated in a way that could potentially contribute to any significant changes to the phylogeographic inference, were among those isolated from patients with another country of origin.

References

31. D. Nguyen, *et al.*, *Journal of Clinical Microbiology* **41**, 2878 (2003).
32. S. Billig, *et al.*, *Scientific Reports* **7**, 6484 (2017).
33. M. A. Behr, W. R. Waters, *The Lancet infectious diseases* **14**, 250 (2014).
34. K. E. Holt, *et al.*, *bioRxiv* (2016).
35. K. I. Bos, *et al.*, *Nature* **514**, 494 (2014).
36. G. L. Kay, *et al.*, *Nat Commun* **6**, 6717 (2015).

Materials and methods

June 18, 2018

Sample collection

To aid population genomic and phylogeographic inferences, we aimed to include large and representative sample collections covering the widest obtainable temporal and geographic range for the study. We included L4 genomes from recent published studies from Argentina (37), Canada (38), Congo-Brazzaville (39), Malawi (40), Netherlands (41), Portugal (42), Russia (43), South Africa (44), Uganda (45), UK (43), USA (46) and Vietnam (47). To further improve temporal structure and resolution we also included three genomes isolated from 18th century Hungarian mummies (48) and genomes from Denmark sampled in the 1960s and 1990s (49). In addition to published genomes, we include 627 previously unpublished genomes from the Americas (Brazil, Peru, Argentina, USA and Canada). Sequencing libraries were prepared as described previously (50) and sequenced on an Illumina platform. Inclusion criteria for individual genomes were as follows: (1) The genome was annotated with a minimum of metadata (sampling year and country of origin), or this information could be obtained from the published articles or by correspondence with personnel involved in the sequencing efforts. (2) The genome had to be L4, as verified by the program TB-profiler (51). Genomes that were determined to be of mixed origin was also excluded (3) If a genome was determined to be a duplicate of an already included isolate, it was not included. (4) All of the following quality control criteria had to be met: (4a) When mapping reads against a H37Rv reference genome, a minimum genome

coverage of 90% had to be reached. (4b) The average read depth across the reference had to exceed 20. We allowed sequence data from different Illumina platforms and with different read lengths. A total of 1,669 genomes passed all inclusion criteria (Dataset S01)

Samples from Argentina and from a study of an Inuit community in Quebec, Canada (38), were mainly from outbreaks and thus harbored limited within-population diversity. In order to control for the effect of densely sampled genomes on certain analyses, we additionally created a down-sampled genome collection, where we allowed just five isolates from each of the Argentinian outbreaks (M and Ra) and ten from the outbreak in Quebec. These genomes were randomly sampled from the full collection by a random number generator. In total, the down-sampled isolate contained 1207 isolates, including the H37Rv reference and an L3 outgroup isolate.

Variant calling

The snippy pipeline v3.1 (52–58) was used for variant calling. Briefly, this entailed mapping reads against H37Rv with the BWA mem algorithm (v.0.7.15-r1110) and marking split hits as secondary, then calling variants with samtools v1.3 and including only reads with a mapping quality of 60 or higher. Variants were then filtered further using FreeBayes (v1.0.2) with a ploidy of 1 and options to exclude (1) alleles if a supporting base quality is less than 20 or the coverage less than 10, (2) alignments if the mapping quality is less than 60, and (3) alleles if the fraction of reads in support of a SNP is not at least 90%. The binomial priors about observation expectations were turned off. The program snpEff v4.11 was then used to annotate SNPs, turning off downstream, upstream, intergenic and 5-prime and 3-prime UTR changes. A whole-genome alignment of all genomes was then built using snippy-core, with a minimum coverage depth of 10 to consider a region part of the core.

The genome sequences from the Hungarian mummies were resolved in a different manner: Sequence read archives from study PRJEB7454 (48) were mapped to H37Rv (59). A minimum read length of 35bp and a minimum mapping quality 30 were imposed. Pilon was run with the following parameters: – variant – mindepth 10 – minmq 30 – minqual 30 and the number of reads found supporting each allele across all variant sites were manually inspected. Three genotypes were inferred from two high coverage sequencing runs (ERR651000 - individual 68 and ERR651004 - individual 92). Genotypes were distinguished based on allele frequencies. For individual 68, genotype 1 was deduced by alleles found between 55-65% and genotype 2 by alleles found between 35-45%. Variants at frequencies of $\geq 95\%$ were considered fixed between the mixed infecting strains. Variants segregating at other frequencies were treated as ambiguous/missing data. For individual 92, we also found evidence indicative of a mixed infection; however, we only felt confident with the major called genotype which comprised $> 90\%$ of reads. All variants at $\geq 95\%$ were called. All other sequencing runs from the project were of low coverage and excluded from the analyses.

We then used an in-house python script (available at: <https://github.com/admiralenola/global4scripts>) to exclude from the alignment SNPs matching any of the following criteria: (1) Located in a known repetitive region (For example PE/PPE genes, annotation file available at github repository), (2) The proportion of ambiguous calls at the locus exceeded 1%, (3) The position was no longer polymorphic after pruning of outbreak isolates (Only applied to the downsampled data set). The final alignment of SNPs consisted of 22,912 sites, with 9,313 of these being parsimony-informative.

Initial phylogenetic analysis

The program Modelfinder (60) as implemented in IQ-TREE was used to infer the optimal substitution model for our genome alignment. A maximum likelihood phylogenetic tree was then

built using the program IQ-TREE v1.4.3 (61), applying the GTR model with 4 gamma categories for rate variation, ascertainment bias on, and 1000 ultrafast bootstrap replicates (62). The L3 isolate SRR1188186 (Quebec, Canada) was used as an outgroup.

AMR gene mutations

Many *M. tuberculosis* AMR mutations are known, but the contribution to the resistance phenotype and penetrance remains unclear for a substantial fraction of variants. We therefore included only high-likelihood AMR mutations relevant for the MDR and XDR definitions. All loci were selected *a priori* for their perceived relevance and strength of association to antimicrobial resistance. For INH we included the *katG* S315T mutation as well as any nonsense or frameshift mutations in the gene plus the classical -15 and -8 *inhA* promoter mutations; For RIF we included all non-synonymous mutations in the resistance-determining region (amino acids 426-450) of *rpoB*; For Kanamycin/amikacin/capreomycin we included *eis* promoter mutations in position -14, -12 and -10 (42, 63) as well as mutation in position 1401 of *rrs*; for fluoroquinolones we restricted the analysis to a manually curated collection of high-likelihood mutations, namely *gyrB* mutations leading to amino acid substitution at positions 461, 499-501 and 642 and *gyrA* mutations resulting in amino acid substitutions at positions 88-94 and 288 (64, 65).

Resistance-associated loci were extracted from the whole-genome alignment using EMBOSS v6.6.0.0 (66) using their H37Rv coordinates. The loci were translated to protein, and sequences sorted by alleles. Mutations were manually annotated on the phylogenetic tree using simple parsimony (*e.g.* an internal node was inferred to have allele A if all descendent nodes had allele A.). The figure was drawn using ITOL (67).

Population genomic inferences

Gene-wise estimates of nucleotide diversity (π), Watterson's theta (θ) and Tajima's D were computed for each gene by parsing the whole-genome alignment into genes using EMBOSS (66) and using the python package Egglib (68) on individual gene alignments. F_{st} values were also calculated using Egglib, specifying groups by (A) country and (B) sub-lineage as identified by TB-profiler. Within each country, we calculated genomic pairwise Hamming distances using the program snp-dists (69).

For each gene we additionally calculated the number of homoplasies and number of homoplastic sites using a Fitch downpass algorithm as implemented in Dendropy (70). Gene-wise parsimony scores were calculated by identifying homoplastic mutations in each gene and summing the Fitch parsimony scores (i.e. the minimum number of independent emergences of each mutation). To investigate potential alterations of fitness induced by the various mutations in and immediately upstream of the *lldD2* gene, we devised a metric for transmissibility associated with each mutation a-kin to (47). First, homoplasies in this gene was categorized as being in the promotor region, in codon 3 and in codon 253. All homoplasy emergences were mapped to the respective branch in the full phylogenetic time tree. For each homoplasy we recorded that subtree's number of descendents, and number of deme transitions. The reasoning behind this was that ancestors with homoplastic mutations increasing transmissibility should have more descendents and more deme transitions per time than ancestors without these mutations. For this latter control group, we randomly extracted subtrees of comparable height distribution to those subtrees with promotor/codon 3/codon 253 mutations, and we labelled this control group as "none". (That is, no homoplastic mutations in *lldD2*). Fig S2 shows a linear model between subtree height and the number of sampled descendents by each mutation category. In order to test whether these categories have different slopes, we used an ANCOVA procedure. First, a simple null model was set up where the number of descendents are dependent on subtree height

but not on the mutation category (including the "none" group). An alternative model is that the relationship between subtree height and the number of descendants vary between these four different mutation categories. For each of these models we weighted the number of deme transitions as $(\text{number of deme transmissions})^2 + 1$, i.e. no transitions got a weight of 1, one transition got a weight of 2, and two transitions got a weight of 5. The ANCOVA analysis rejected the null model, showing significant preference for the per-group model (F-test: $p=0.038$). Analysis of individual height:group interaction terms showed that the coefficient for promotor mutations were significantly different from zero, indicating a positive association between *ltdD2* promotor mutations and transmissibility. Note that if the weighting by deme transitions is removed, the ANCOVA analysis no longer significantly prefers the alternative model (F-test: $p=0.114$), and there is no evidence for homoplasy group interaction with height and number of descendants.

Phylogenetic and Phylogeographic inference

In order to estimate substitution rates by means of sampling-date-calibrated Bayesian evolutionary analyses (71), we downsampled the collection to a manageable size by including a maximum of 20 randomly chosen isolates from each country. This resulted in a sample collection of 269 genomes. Importantly, three ancient *M. tuberculosis* genomes from 18th century hungarian mummies (48) as well as five Danish isolates from the 1960s (49) were included to provide temporal structure to the data.

A total of 7,994 variable sites remained after selecting the 269 genomes. A preliminary check using TempEst (72) confirmed a moderate but highly significant temporal signal in the data (Fig .S4), which was also confirmed by tip-randomization (see below). A GTR substitution model was chosen based on model testing as described above. Based on marginal likelihood estimation (MLE), an exponential demographic model was found to best fit the data (tested against the constant population size and GMRF skyride models (73)). Also based on MLE, an

uncorrelated relaxed clock was favored over a strict clock model. Three independent BEAST MCMC chains were run and convergence to a stationary posterior distribution was confirmed both within and between chains. These analyses resulted in an estimated substitution rate of 4.84×10^{-8} (95% HPD: 4.16×10^{-8} - 5.44×10^{-8}) substitutions per site per year and an estimated time of the MRCA in 1096 CE (95% HPD: 955-1231). To assess the robustness of the temporal inference, we performed 10 additional runs after randomization of the sampling dates (74). None of the randomized runs had rate estimates overlapping with the inference using real sampling dates (Fig. S3), supporting the robustness of the original inference.

We then ran a larger dataset (1,207 genomes, after retaining a maximum of five representatives from each of three densely sampled outbreaks from Argentina and ten from Quebec, Canada) in Beast 1.8.4 (75) with a fixed substitution rate as estimated above. The MRCA of this tree was inferred to have existed in the year 1157 (HPD intervals not reported due to our application of a fixed rate in this analysis). The overall accuracy of the dating inferences in Beast was further supported by independent analyses applying LSD v0.3 (76), resulting in an estimated TMRCA in 1195 CE (95% HPD: 1061 - 1270) for the 1,207 genome dataset. The tree generated for 1207 genomes in BEAST 1.8.4 was used as input tree for phylogeographic analyses employing both the BASTA module in BEAST2 and simple discrete trait analyses (DTA) as implemented in BEAST 1.8.4 (?).

In order to reduce computational complexity and increase post-analysis interpretability we collapsed country of origin information into five distinct UN regions: North America (USA and Canada), South America (Brazil, Argentina, and Peru), Africa (DRC, Malawi, South Africa and Uganda), Europe (Denmark, the Netherlands, Hungary, Portugal, Russia, and UK) and South East Asia (Vietnam).

In the discrete trait model (77), we used a GTR model, restricted the clock rate to 4.84×10^{-8} , and set the starting tree as specified above. The population size prior was set as constant. A

symmetric deme substitution model was used and we turned Bayesian stochastic search variable selection on. Ancestral states were reconstructed at all nodes. Other priors were left at default values. The chain was run for 10,000,000 generations with logging every 10,000th iteration, and three such runs were combined to create the final posterior sample of trees and parameters. We verified chain convergence and good mixing and an ESS > 200 for all parameters using Tracer (78). A maximum clade credibility (MCC) tree was created using TreeAnnotator (<http://beast.community/treeannotator>), with 20% of the chain discarded as burn-in. The resulting tree displayed a European root with 92% posterior probability.

A separate phylogeographic reconstruction of the L4.3 RdRio family was also performed. This was completed by manually extracting the RdRio sub-tree from the full time tree and then setting up a new DTA run consisting of these 243 isolates alone. For this analysis we acquired patient country of origin data for genomes from our Portuguese and Dutch collections, which led to a changed country of origin for 13 genomes in this data set. To restrict the number of possible demes of low/single sample size we collapsed countries into eight different geographic categories: Iberia (Portugal, Spain, Cape Verde), North Europe (UK, Russia, Bosnia, Netherlands, Germany), Peru, Atlantic South America (Venezuela, Brazil, Aruba, Suriname), Congo/Angola, Malawi/Mozambique, Uganda and South Africa. The population size prior was set to 10,000. Other than this, the analysis was run with the same parameters as for the full data set DTA analysis. The rationale for placing Cape Verde in the Iberia category is that Cape Verde was uninhabited prior to Portuguese settlement in the 15th century.

As a complement to DTA, we used the BEAST2 module BASTA v2.3.1 (79) for phylogeographic inferences. We specified a migration model with the same five demes as above. The initial values for deme transition rate was set to 1.0×10^{-3} and sub-population to 6,000, these numbers corresponding to the median outputs from DTA. The rate matrix and population size priors were given log-normal prior distributions with M=-10 and S=2.0 and M=9.0 and S=0.6,

respectively. Since it was not possible to place deme restrictions on internal nodes in BASTA, we artificially introduced an isolate with branch length 1.0×10^{-10} from the root, and the location of this isolate was set to correspond to each of the five demes in different runs. The results of these five runs were subsequently evaluated jointly. We ran each chain for 1,000,000 generations with storing set to every 1,000th iteration. We disabled all scaling operators except the rate scaler, which was given a scale factor of 0.8 and a weight of 1.0, and the population size scaler which was given a weight of 3 and a scale factor of 0.8 and degrees of freedom set to 1 (Sub-population sizes effectively set to equal). Since we knew that most recent common ancestor (MRCA) of all isolates existed roughly around the year 1100 CE, prior to European colonization of the Americas, we discarded all trees with a root inferred to be from North or South America. The parameter logs were inspected in the same way as described for DTA, and an MCC tree made using TreeAnnotator v.2.4.5, with burn-in set to 20 % and node heights set to median. All BEAST and BEAST2 runs were run locally or on the CIPRES science gateway (80).

Migration matrices (81) were constructed to visualize the overall patterns of migration inferred by the two methods. Both methods inferred Europe to have played a pivotal role in sourcing the rest of the world with TB (Fig. S5).

In order to study migration over time, conceptually mirroring the methodology used in (81), we wrote an in-house script (available at <https://github.com/admiralenola/global4scripts>) to read the MCC tree from the BASTA runs using the ETE toolkit (82) and traversed the time tree in a sliding window fashion, for each year writing out the number of branches corresponding to the five different demes. In our analyses, migration events were set to occur on nodes, but could in reality have occurred at any point along the branch downstream of this node. This introduces a slight bias towards inflated ages of migration events, which is most pronounced for very early migration events but negligible for later migrations due to extensive branching.

References

37. V. Eldholm, *et al.*, *Nature Communications* **6** (2015).
38. R. S. Lee, *et al.*, *Proceedings of the National Academy of Sciences* **112**, 13609 (2015).
39. S. Malm, *et al.*, *Emerging Infectious Disease journal* **23**, 423 (2017).
40. J. R. Glynn, *et al.*, *PLoS ONE* **10**, e0132840 (2015).
41. J. M. Bryant, *et al.*, *BMC Infectious Diseases* **13**, 110 (2013).
42. J. Perdigao, *et al.*, *BMC Genomics* **15**, 991 (2014).
43. N. Casali, *et al.*, *Nat Genet* **46**, 279 (2014).
44. K. A. Cohen, *et al.*, *PLoS Med* **12**, e1001880 (2015).
45. A. L. Manson, *et al.*, *Nat Genet* **49**, 395 (2017).
46. T. S. Brown, *et al.*, *BMC Genomics* **17**, 947 (2016).
47. K. E. Holt, *et al.*, *bioRxiv* (2016).
48. G. L. Kay, *et al.*, *Nat Commun* **6**, 6717 (2015).
49. T. Lillebaek, *et al.*, *Int J Med Microbiol DOI:10.1016/j.ijmm.2016.05.017* (2016).
50. V. Eldholm, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **113** (2016).
51. F. Coll, *et al.*, *Genome Medicine* **7**, 51 (2015).
52. T. Seeman, Snippy (2015).

- 53. H. Li, *et al.*, *Bioinformatics* **25** (2009).
- 54. H. Li, *arXiv* (2013).
- 55. E. Garrison, G. Marth, *arXiv* (2012).
- 56. E. Garrison, Vcflib: A C++ library for parsing and manipulating VCF files (2012).
- 57. P. Danecek, *et al.*, *Bioinformatics* **27**, 2156 (2011).
- 58. P. Cingolani, *et al.*, *Fly* **6**, 80 (2012).
- 59. S. T. Cole, *et al.*, *Nature* **393**, 537 (1998).
- 60. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermini, *Nat Meth* **14**, 587 (2017).
- 61. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, *Molecular Biology and Evolution* **32**, 268 (2015).
- 62. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, *Molecular Biology and Evolution* **30**, 1188 (2013).
- 63. S. B. Georgiou, *et al.*, *PLoS ONE* **7**, e33275 (2012).
- 64. M. R. Farhat, *et al.*, *Journal of Clinical Microbiology* **54**, 727 (2016).
- 65. S. Malik, M. Willby, D. Sikes, O. V. Tsodikov, J. E. Posey, *PLoS ONE* **7**, e39754 (2012).
- 66. P. L. Rice, A. Bleasby, *Trends in Genetics* **16**, 276 (2000).
- 67. I. Letunic, P. Bork, *Nucleic Acids Research* **39**, W475 (2011).
- 68. S. De Mita, M. Siol, *BMC Genetics* **13**, 27 (2012).

69. T. Seeman, SNP-dists (2017).
70. J. Sukumaran, M. T. Holder, *Bioinformatics* **26**, 1569 (2010).
71. A. J. Drummond, G. Nicholls, A. Rodrigo, W. Solomon, *Genetics* **161**, 1307 (2002).
72. A. Rambaut, T. T. Lam, L. M. Carvalho, O. G. Pybus, *Virus Evolution* **2** (2016).
73. V. N. Minin, E. W. Bloomquist, M. A. Suchard, *Molecular Biology and Evolution* **25**, 1459 (2008).
74. A. Rieux, C. Khatchikian, TipDatingBeast: Using Tip Dates with Phylogenetic Trees in BEAST (R package) (2015).
75. A. Drummond, A. Rambaut, *BMC Evol Biol* **7**, 214 (2007).
76. T.-H. To, M. Jung, S. Lycett, O. Gascuel, *Systematic Biology* **65**, 82 (2016).
77. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, *PLoS Comput Biol* **5**, e1000520 (2009).
78. A. Rambaut, A. J. Drummond (2003).
79. N. De Maio, C.-H. Wu, K. M. O'Reilly, D. Wilson, *PLOS Genetics* **11**, e1005421 (2015).
80. M. Miller, W. Pfeiffer, T. Schwartz, *Proceedings of the Gateway Computing Environments Workshop* pp. 1–8 (2010).
81. M. B. O'Neill, *et al.*, *bioRxiv* (2017).
82. J. Huerta-Cepas, F. Serra, P. Bork, *Molecular Biology and Evolution* **33**, 1635 (2016).

Supplementary Figures

June 18, 2018

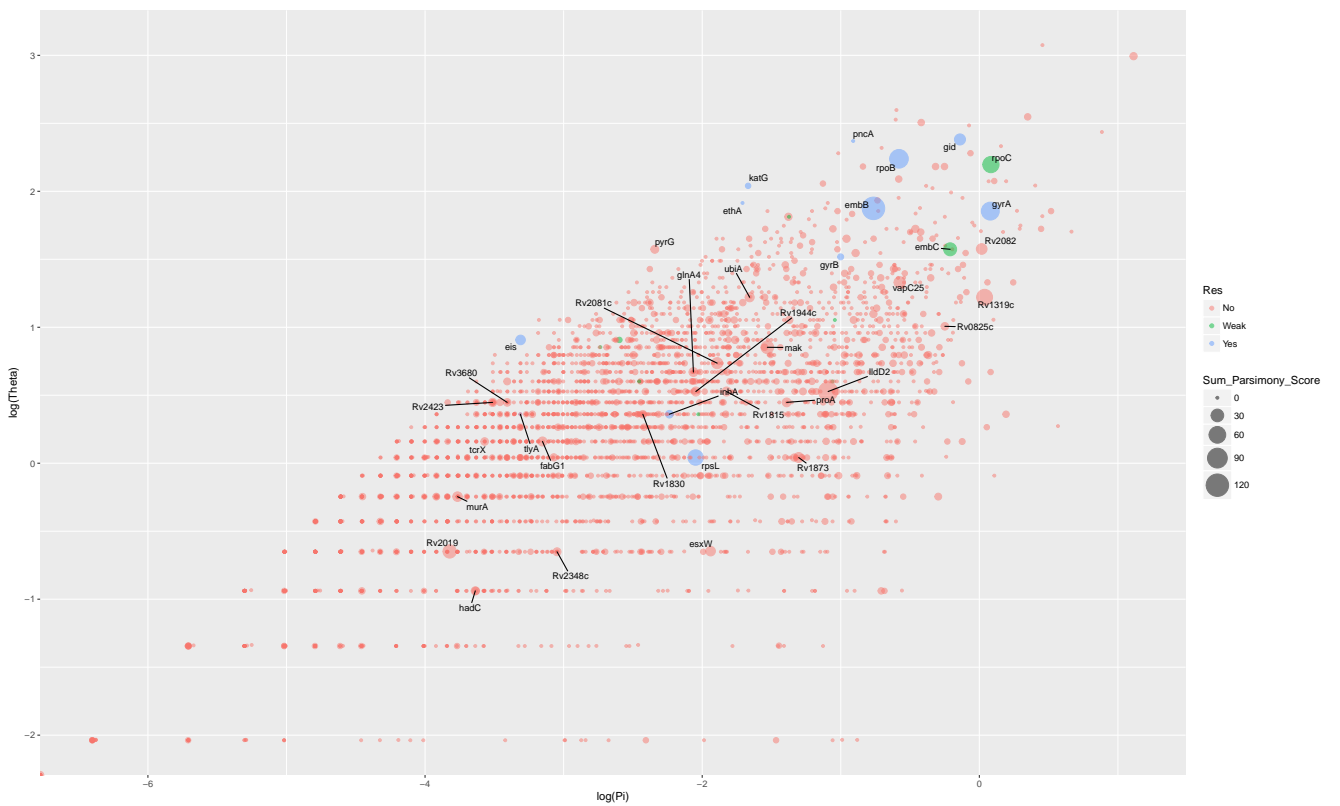


Figure S1: *Genome-wide assessment of homoplastic mutations, presented as gene-wise parsimony scores as a function of θ and nucleotide diversity (π).*

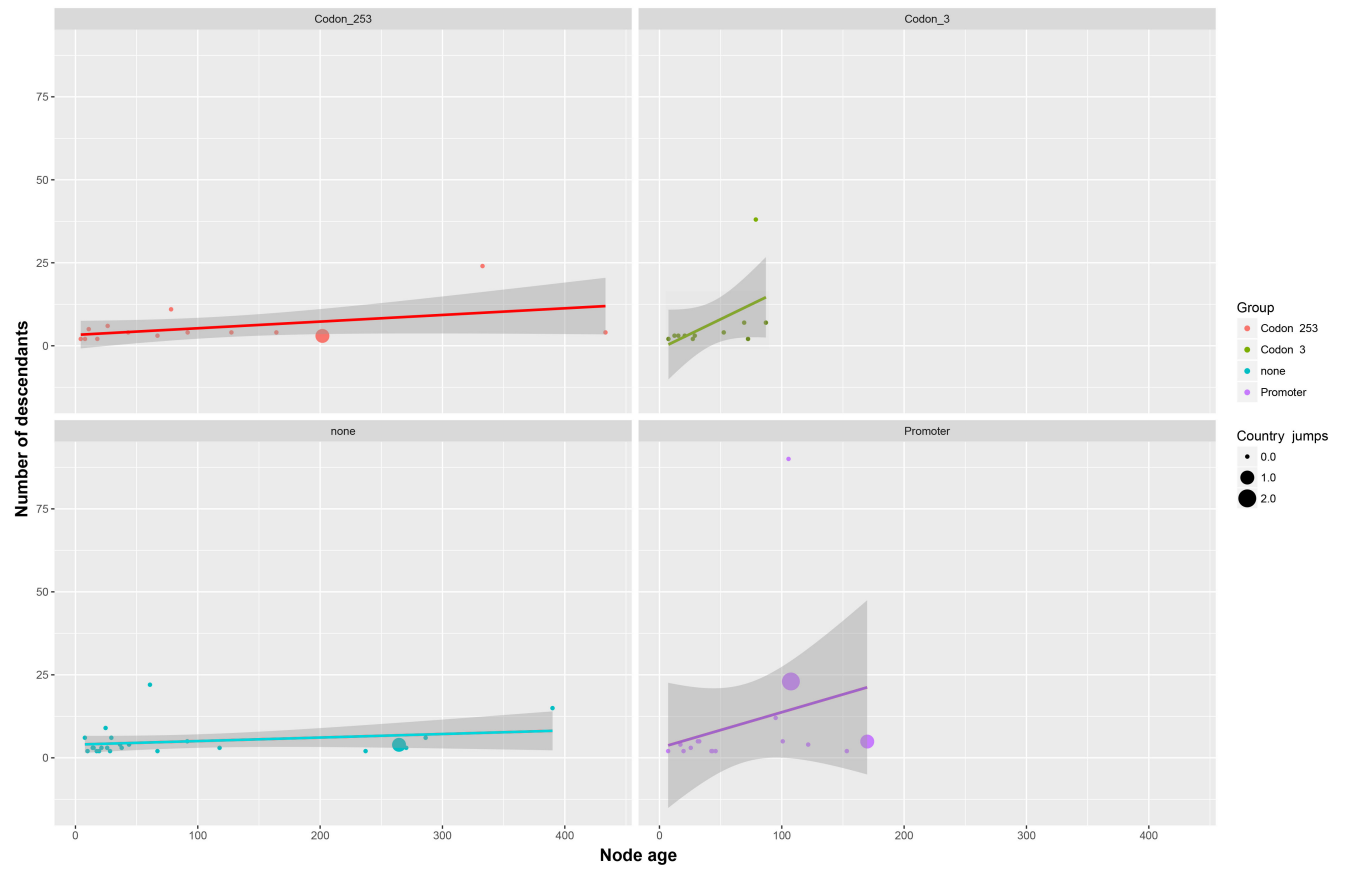


Figure S2: Node descendants as a function of node age for clones harboring different groups of *lldD2* mutations. For each group, a linear regression line including 95% confidence intervals was fitted to the data. Analysis of variance between the four groups did not identify significance difference between them ($p=0.08$), but when the number of country jumps was used to weight the observations, the slope of the groups were found to be significantly different ($p=0.04$)

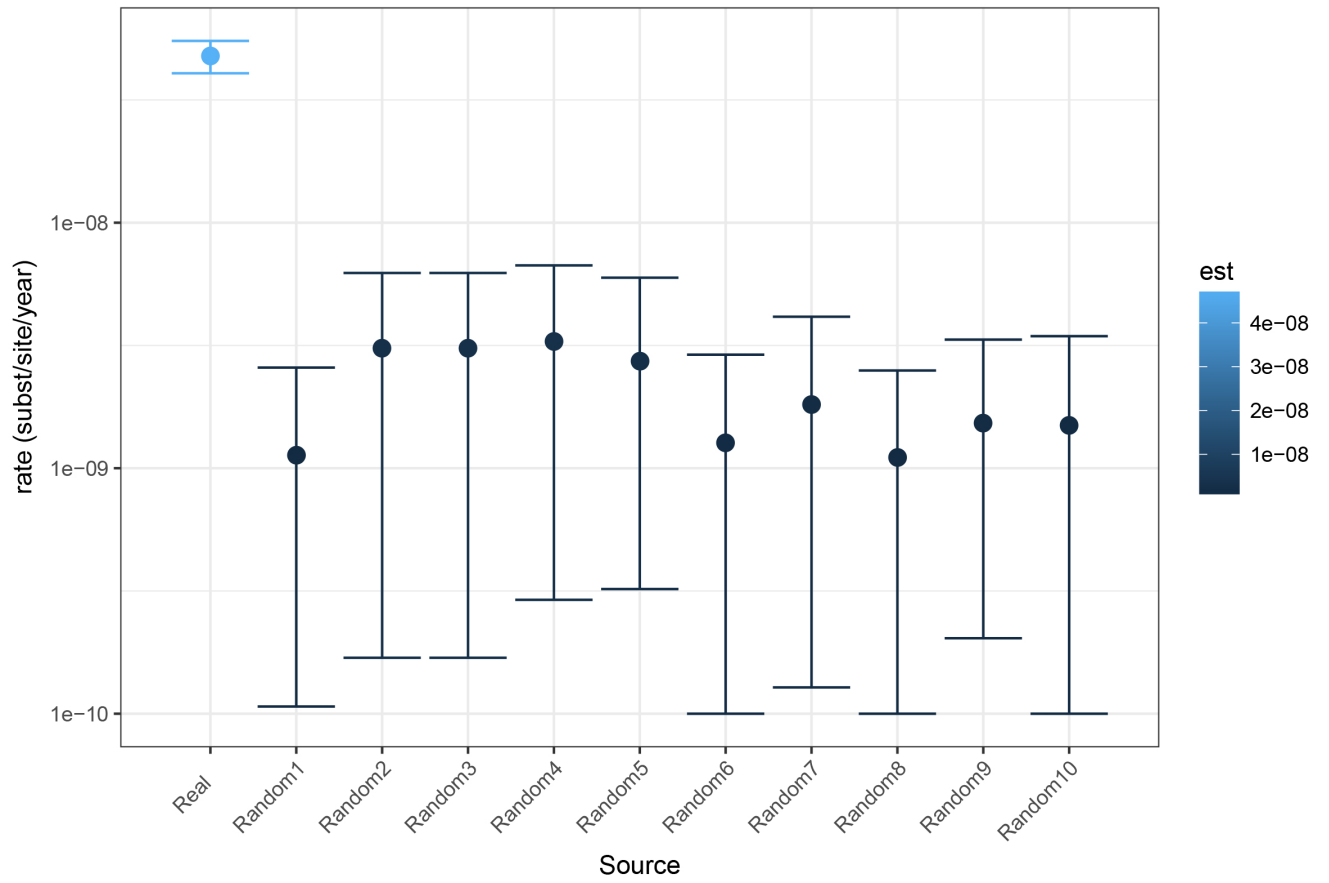


Figure S3: *Tip-randomization results*

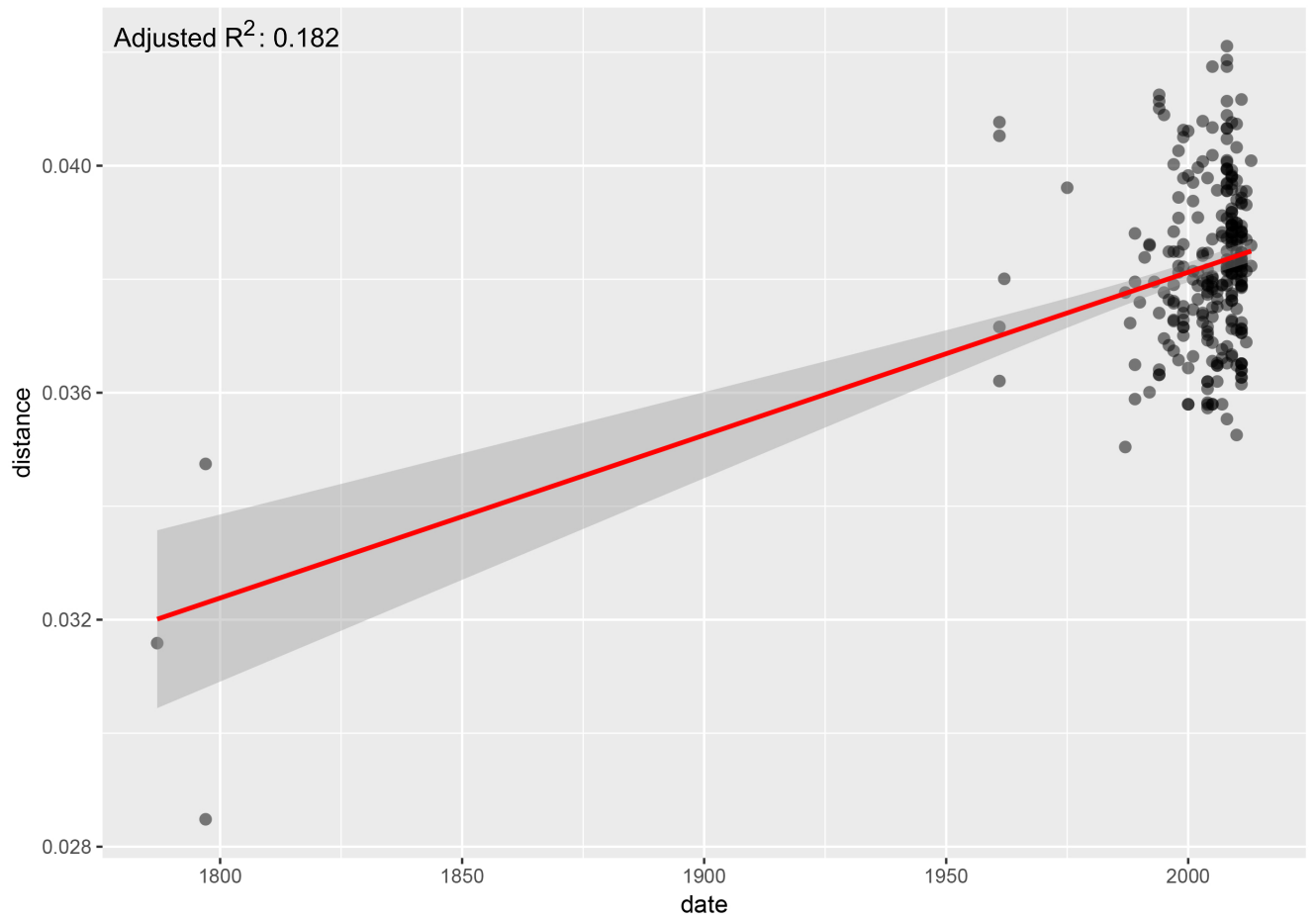


Figure S4: Root-to-tip analysis performed on a global down-sampled collection containing 269 genomes

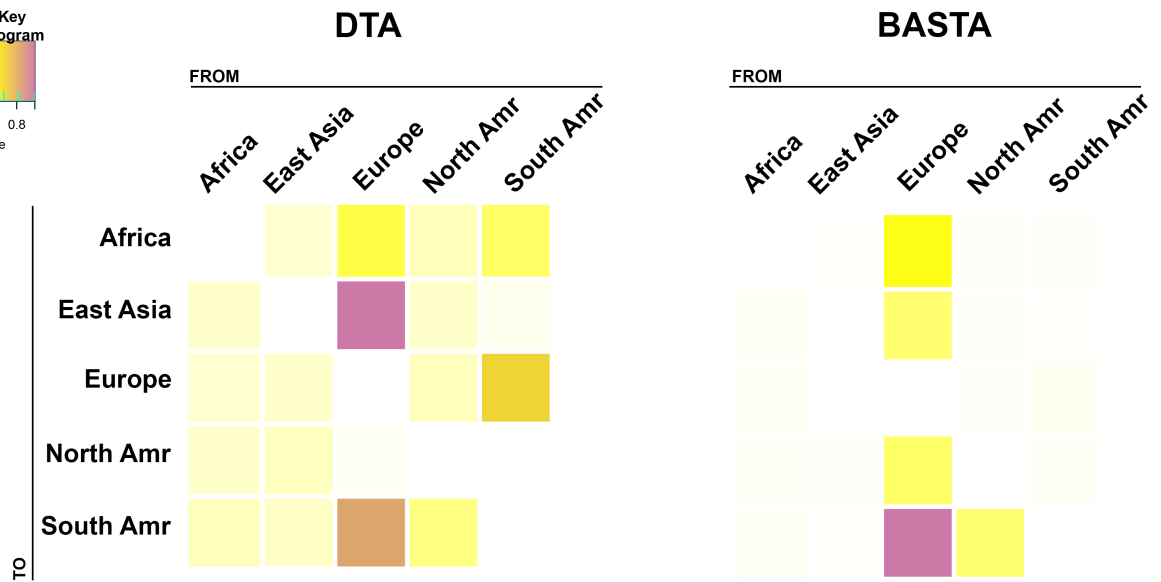
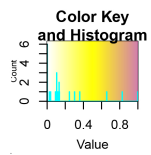


Figure S5: Migration matrices inferred with DTA and BASTA visualized as heat maps.

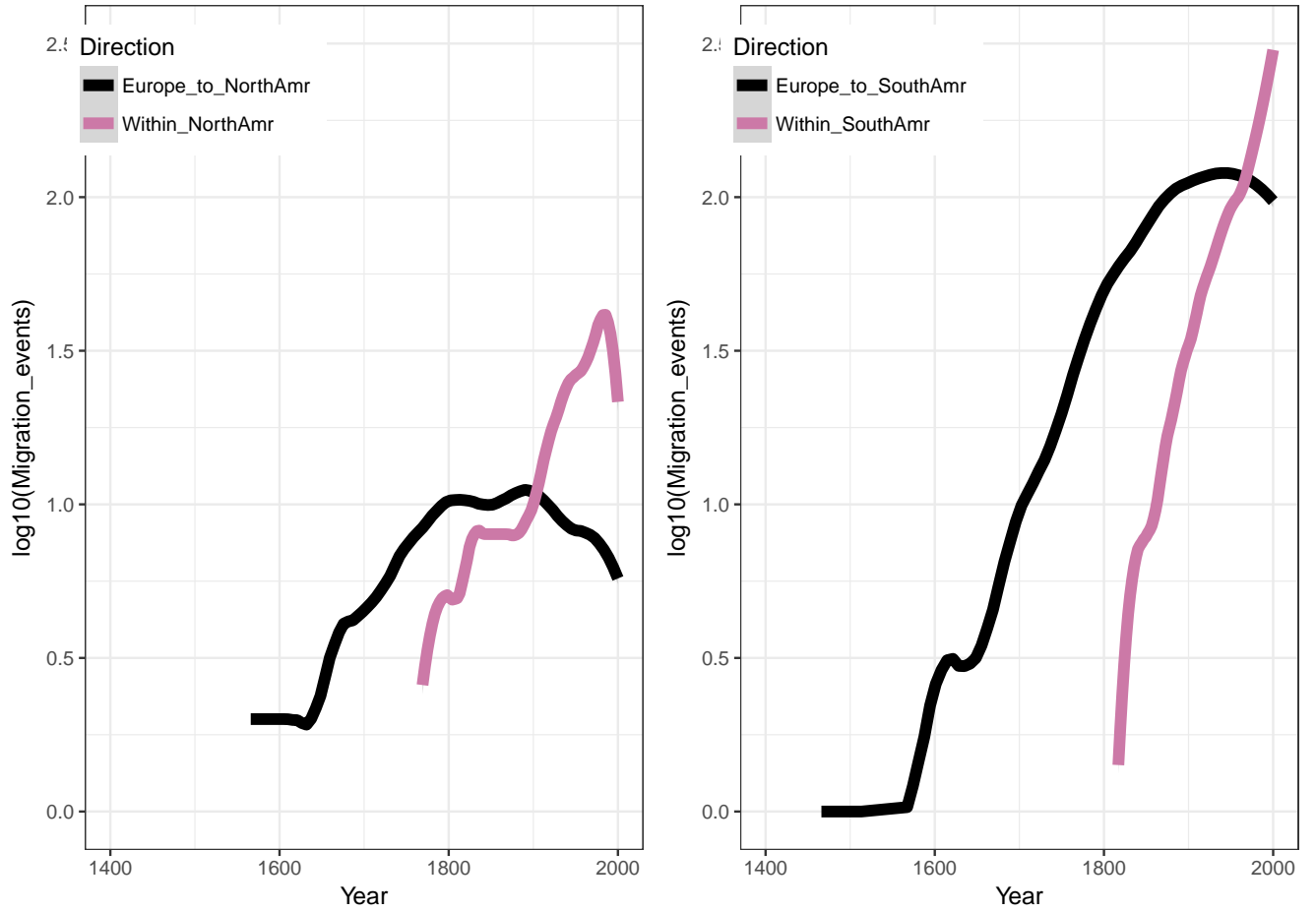


Figure S6: *Inferred migration of L4 over time from Europe to North and South America, as well as within the continents, employing BASTA.*

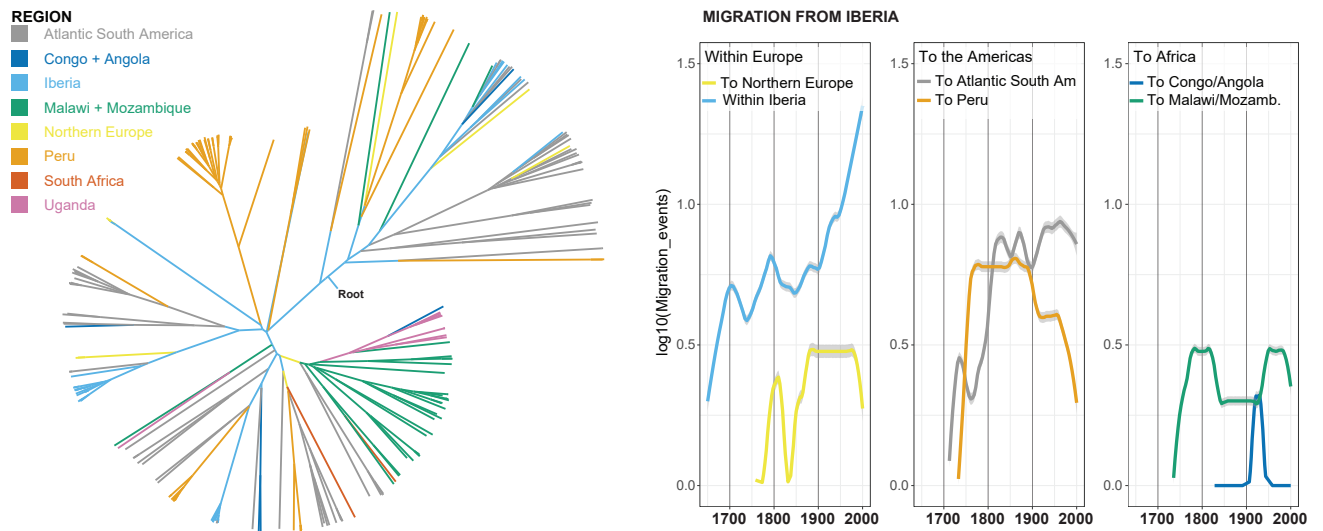


Figure S7: *Phylogeographic reconstruction of the RdRio family. In the left panel, branches are colored according to inferred location based on discrete trait analysis in Beast. The right panel summarizes the intensity of migration over time from Iberia (the inferred origin).*

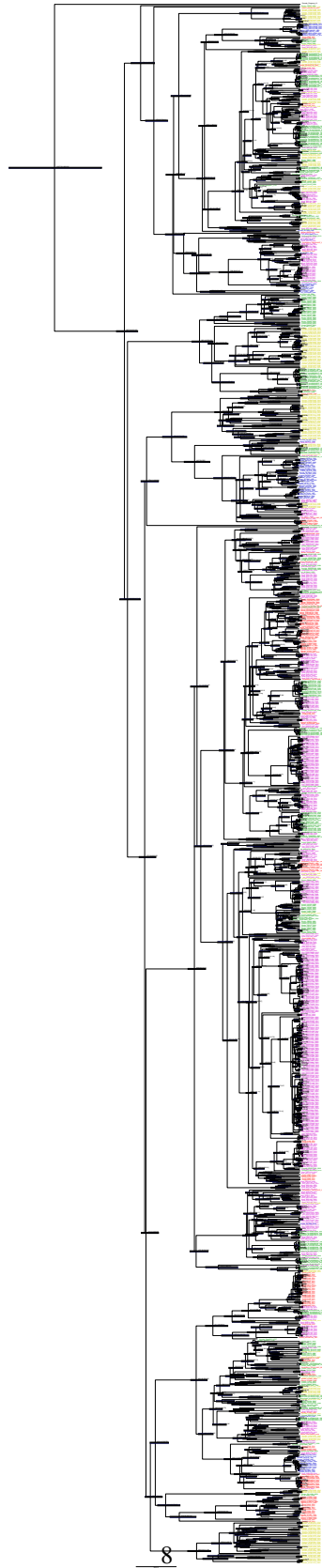


Figure S8: *Full temporal phylogeny of L4 including node age 95% HPD intervals.*

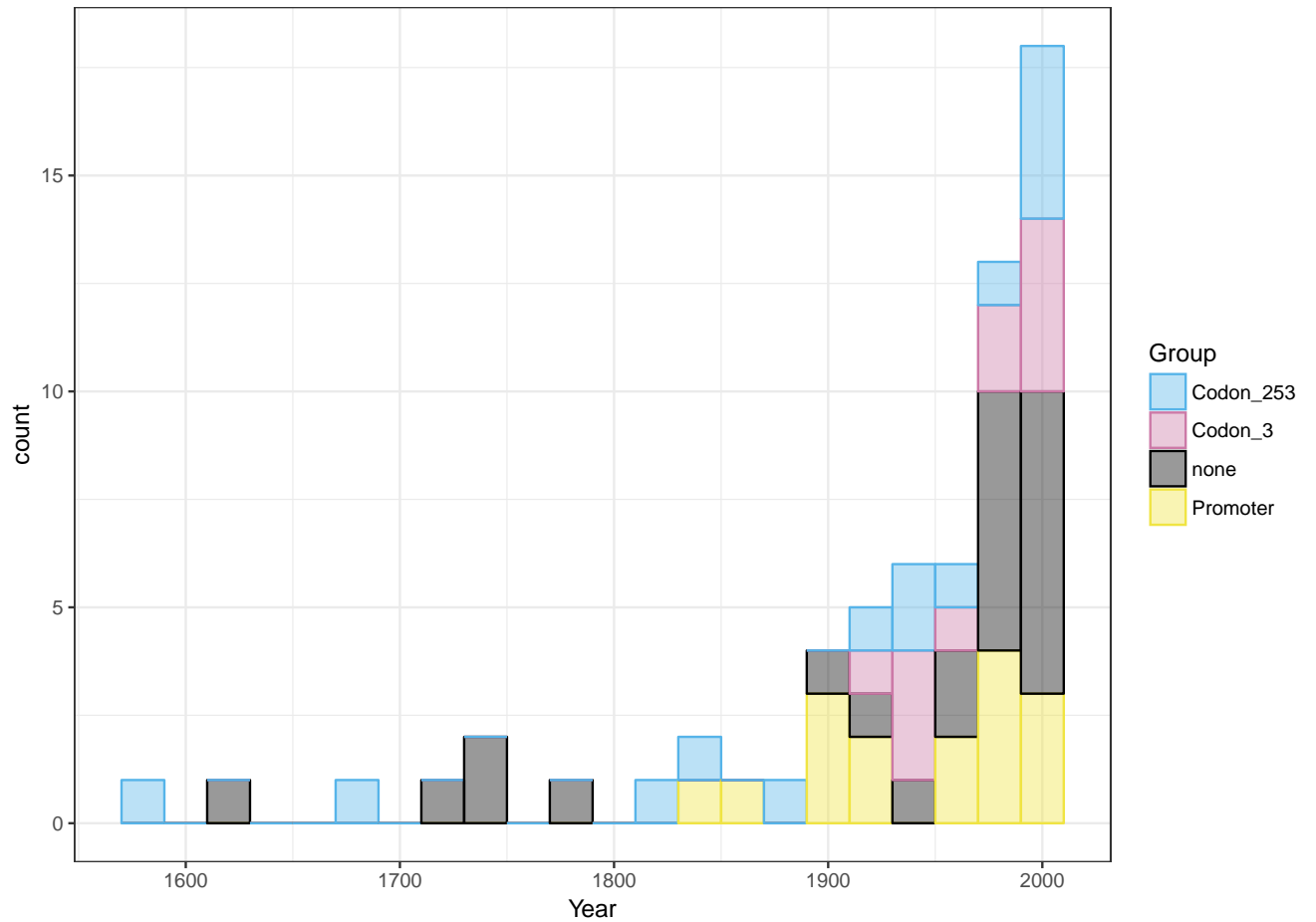


Figure S9: Histogram summarizing the emergence of *lldD2* mutations over time. As a reference, a control is included containing nodes where such mutations did not emerge. Analysis of individual height:group interaction terms showed that the coefficient for promotor mutations were significantly different from zero, indicating a positive association between *lldD2* promotor mutations and transmissibility. Note that if the weighting by deme transitions is removed, the ANCOVA analysis no longer identifies any significant differences between the groups (F -test: $p=0.114$)